Openeuler用户画像标签体系建模：

**目的：促进用户的活跃度，做最活跃的社区。**

**原始数据：**

* Issue\_info：title、creator、body、label
* Comment\_info：commenter、body
* Event\_info：actor、event\_type

**事实标签：**

* 社区历史趋势：Issue、PR创建数量、评论数量
* 目前已有标签的出现趋势
* 用户的行为趋势、评论趋势、创建issue pr的趋势

**模型标签：**

* **活跃度**：根据评论数、PR数综合计算活跃程度
* **专业偏好**：根据用户参与Issue的label进行评估
* **社会属性**：根据用户参与其他用户的Issue进行计算
* **用户流失概率**：综合用户各方面表现进行预测
* **用户活跃时间偏好**：用户活跃的时间段
* **用户-用户关系图：**
  + **ForceAtlas2算法：**
    - 是为力引导布局，模仿物理世界的引力和斥力，自动布局直到力平衡

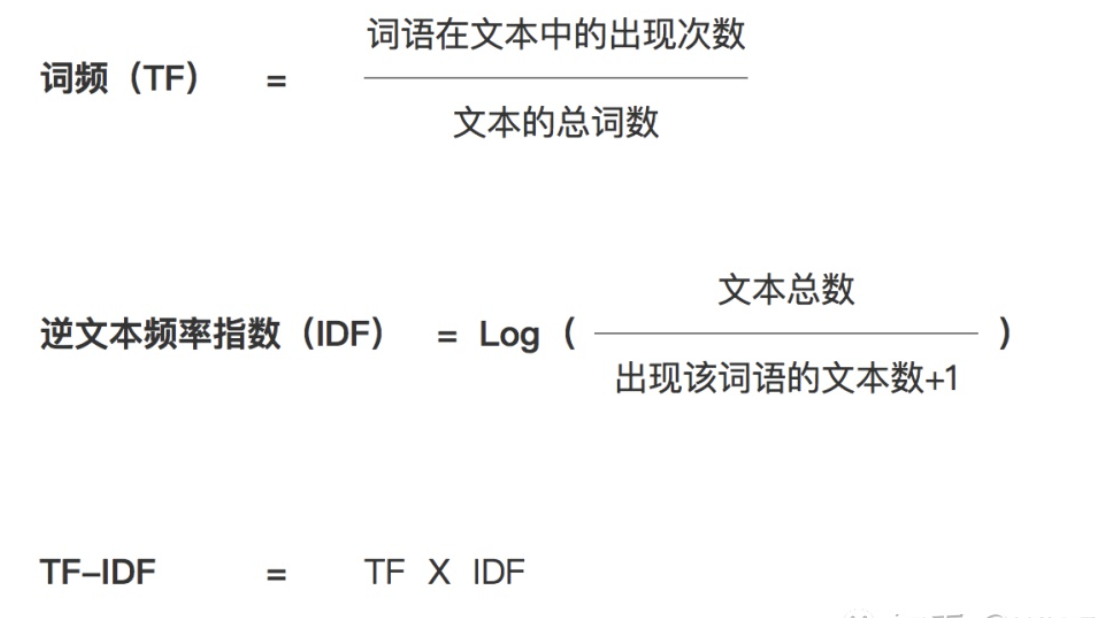


* + **Fruchterman Reingold算法：**
    - 算法通过考虑原子间引力和斥力的互相作用，计算得到节点的速度和加速度。依照类似原子或者行星的运动规律，系统最终进入一种动态平衡状态。

**预测标签：**

* 基于规则的方法：
  + - 优点：可解释性强、匹配速度快
    - 缺点：需要人工定义规则、存在误判情况、模型较难适应复杂场景
* 基于主题模型的方法：
  + 1.1无监督方法（LDA、TF-IDF、word2vec）：
    - 优点：无需人工标注信息
    - 缺点：主题类别需要人工确定、精确度低
  + 1.2典型算法

1) TF-IDF提取关键词：



原理：**某个词在文章中的TF-IDF越大，那么一般而言这个词在这篇文章的重要性会越高，所以通过计算文章中各个词的TF-IDF，由大到小排序，排在最前面的几个词，就是该文章的关键词。优点是简单快速，而且容易理解。缺点是有时候用词频来衡量文章中的一个词的重要性不够全面，有时候重要的词出现的可能不够多，而且这种计算无法体现位置信息，无法体现词在上下文的重要性。**

2) LDA（Latent Dirichlet Allocation）提取主题：

LDA模型包含**包含词、主题和文档三层结构**

* + - 实现方法：Genism开源包
  + 2.1有监督方法（SVM、k-means、朴素贝叶斯）：

将文本进行处理后，转换成向量集的形式再使用文本分类算法。

* + - 优点：精确度较高、部署较简单
    - 缺点：数据需要标注、训练速度较慢
  + 2.2 典型算法

1. SVM(Joachims T. Making large-scale SVM learning practical[R]. Technical report, 1998.)

给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。

实现方法：Python Scikit-learn包

1. K-means(Krishna K, Murty M N. Genetic K-means algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1999, 29(3): 433-439.
2. 朴素贝叶斯(Rish I. An empirical study of the naive Bayes classifier[C]//IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001, 3(22): 41-46.)
   * 3.1 强化学习方法（Bert、TextCNN、TextRNN）：
     + 优点：精确度高、模型复杂可以应对多种场景、部署较复杂
     + 缺点：数据需要标注、训练速度慢
   * 3.2 典型算法

1) Bert(Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.)基于Transformer算法的双向编码表征算法

2) TextCNN(Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.)利用卷积神经网络对文本进行分类的算法

3) TextRNN(Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015)利用RNN循环神经网络解决文本分类问题